The Curse of Recursion: Training on Generated Data Makes Models Forget

Year: 2023 | Citations: 375 | Authors: Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Y. Gal, Nicolas Papernot

---

Abstract

Stable Diffusion revolutionised image creation from descriptive text. GPT-2, GPT-3(.5) and GPT-4 demonstrated astonishing performance across a variety of language tasks. ChatGPT introduced such language models to the general public. It is now clear that large language models (LLMs) are here to stay, and will bring about drastic change in the whole ecosystem of online text and images. In this paper we consider what the future might hold. What will happen to GPT-{n} once LLMs contribute much of the language found online? We find that use of model-generated content in training causes irreversible defects in the resulting models, where tails of the original content distribution disappear. We refer to this effect as Model Collapse and show that it can occur in Variational Autoencoders, Gaussian Mixture Models and LLMs. We build theoretical intuition behind the phenomenon and portray its ubiquity amongst all learned generative models. We demonstrate that it has to be taken seriously if we are to sustain the benefits of training from large-scale data scraped from the web. Indeed, the value of data collected about genuine human interactions with systems will be increasingly valuable in the presence of content generated by LLMs in data crawled from the Internet.